Towards Automated Analytics of Research Publications

ESIP Winter 2022
January 18-21, 2022
Irina.Gerasimov@nasa.gov

https://disc.gsfc.nasa.gov/ NASA/Goddard Earth Sciences Data and Information Services Center (GES DISC)

Irina Gerasimov[1,2], Jim Acker[1,2], Andrey Savtchenko[1,2], Jerome Alfred[1,2], Armin Mehrabian[1,2], Binita KC[1,2], and Jennifer Wei[1]
[1]Code 610.2, NASA Goddard Space Flight Center, Greenbelt, MD, USA [2]ADNET Systems Inc., Lanham, MD, USA

Enabling Open Science

Abstract: *For readers of scientific publications it remains a big challenge to unambiguously relate the published research with the data used. To a substantial degree it is attributed to authors, journals, editors, and reviewers not prioritizing correct data citation, which impacts traceability, repeatability, and giving credits to published authors and their funding sources. Furthermore, uniform classification of the content of the published research is hampered by journals using journal specific topics and letting authors to assign free text keywords to their papers. We demonstrate automated analytics methods for extracting and relating datasets used and the research application areas by processing 1,300 research papers that referenced the NASA Giovanni service (but probably not the datasets in particular) as supporting their publication process.*

Our automated analysis of the publication texts helps the reader of the paper to determine:

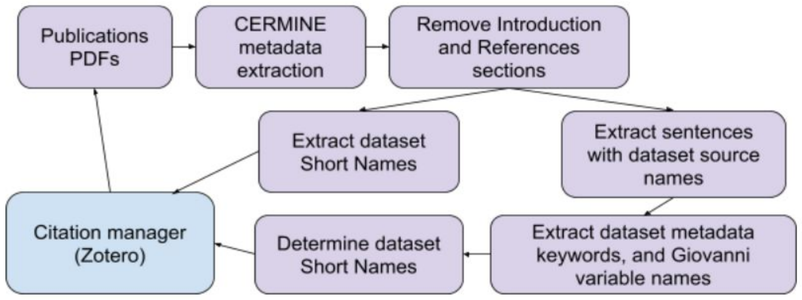
- The application area of the publication
- The data sources (names of missions, instruments and re-analysis models) used to study this application area
- (Some of the) datasets and variables used in research

This information opens the possibility of tracing and reproducing research results reported in the paper and evaluating scientific and community impact of the datasets used in research.

Analytics of Research Publications that use NASA Giovanni

- NASA Giovanni Visualization and Analysis system has been available since 2000. It enables users to perform various analyses using many popular Earth Science datasets from various Earth observation missions and re-analysis models.
- Researchers began citing Giovanni in 2004, and since 2013 at least 200 journal articles and book chapters citing it have been produced every year. At the beginning of 2022 the total number of publications that used the Giovanni system exceeded **2,600**.
- GES DISC staff traces Giovanni-citing research articles to determine the impact of NASA Giovanni and the datasets through the following metrics:
  - NASA Giovanni parameters
  - Data sources: instruments and models
  - Applied research areas
- Each month 20-30 articles are collected through Google Scholar keyword search. These articles are manually reviewed and the results are not uniform. If the review style/parameters need to be changed later then the procedure must be applied to all previously reviewed papers.
- Automation of the paper review process saves the review time and enables uniform paper processing that can be adjusted depending on the desired analytic parameters.

Natural Language Processing Pipeline

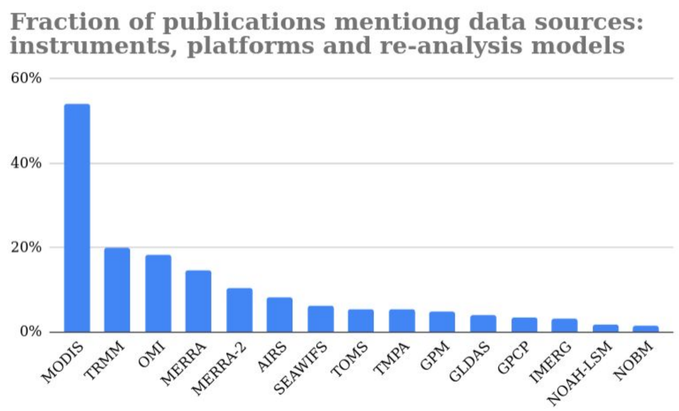


- All collected citations are stored in a citation management system, Zotero, which manages citations and their PDF attachments.
- Zotero Python API interface allows users to link PDF attachments and citation metadata.
- Java CERMINE package is used to convert PDF to ASCII and split publication text into the sections.
- Dataset metadata keywords, mission, instrument and model names, and Giovanni variable names are extracted from the preprocessed publication text sentences, and arranged for further analysis.

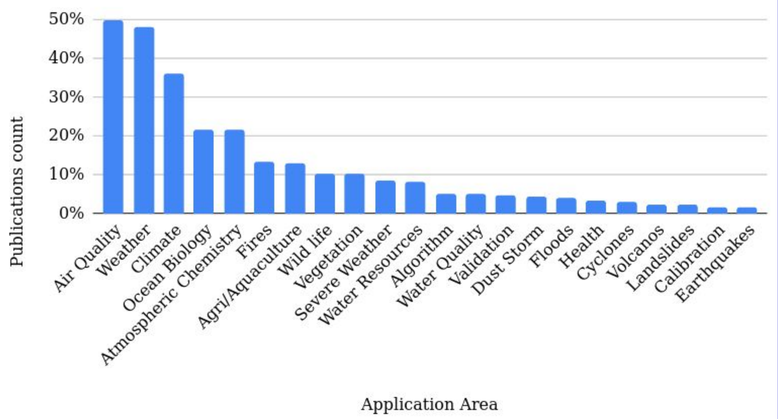
Attributing Publications to the Data Sources and Datasets

The following statistics were collected after ~1300 publication PDF files covering 2016 - 2021 were processed:

- **1%** of papers used DOIs to cite datasets.
- **17%** of papers used dataset Short Names (unique string identifiers of individual NASA Earth Science datasets; dataset name can be mapped to the dataset DOI) to identify datasets.
- *Observation: even if the paper used DOIs or Short Names to identify some of the datasets, it does not necessarily reference these attributes to identify other datasets used in the paper.*
- **87%** of papers mentioned data source such as platform and/or instrument or reanalysis model name.
- **13%** of papers did not indicate datasets or sources of the data but listed NASA Giovanni variable names such as “Sea Surface Temperatures”, “Wind”, “Chlorophyll a”, etc. acquired from Giovanni. In most of these cases it is possible to determine a dataset or a group of the datasets based on the stated variable name, as well as temporal and spatial resolution.



NASA Giovanni Application Areas



- The application area names (topics) were developed in collaboration with a ESIP Usage Based Discovery project (Chris Lynnes). GES DISC scientists modified the list of topics, adding the new and combining the original ones.
- Each topic contains the list of terms that, when appearing in the publication title or abstract, indicate that the paper is *likely* about that topic.
- The paper’s titles and abstracts were processed to extract topic terms based on which the paper was assigned to one or more topics without prioritizing.
- The prevailing research topics studied using Giovanni service are: Air Quality, Weather, Climate, Ocean Biology, Atmospheric Chemistry, Fires, Agri/Aquaculture, Wildlife and Vegetation.

Topic	Terms
Air Quality	aerosol,air quality,air pollution,pm 2.5, pm 10,black carbon,no2,nox,so2,sulphur dioxide,sulphate,acid,carbon monoxide,haze,dust,smoke
Weather	weather, precipitation,rain,snow,humidity,wind chill,cloud,water vapor,air temperature, heat wave
Climate	climate,global warming,global change, ENSO, QBO
Ocean Biology	plankton,chlorophyll,macroalga,algae,barnacle,coral,coccolithophore,cyanobacteria,diatom,sea cucumber,crab,whale,fish,diazotroph,shrimp,squid,shell,oyster
Atmospheric Chemistry	co2,carbon dioxide,ch4,methane,o3,ozone,no2,nitro,so2,sulphur dioxide
Fires	wildfire,fire,burn,smoke
Agri/Aquaculture	agriculture,food,crop,farm,aquaculture,fishery
Animals	bird,woodcock,petrel,rodent,sea cucumber, crab,whale,fish,diazotroph,shrimp,squid,shell,oyster,reproductive,breed,habitat
Vegetation	vegetation,ndvi,primary product
Severe Weather	severe weather,extreme heat,storm,tornado,lightning
Water Resources	drought,runoff,run-off,water resource
Water Quality	water quality,algal,algae,coral bleach,eutrophication,nutrient loading,nutrient pollution
Earthquakes	earthquake,seismic,tremor
Dust Storm	dust storm,sandstorm
Floods	flood,inundation,jokulhlaup,glof
Cyclones	hurricane,cyclone,typhoon
Health	cholera,uv index,ultraviolet,vitamin d,zoonotic
Volcanos	volcan,lava,tephra,tuff,pyroclastic
Landslides	landslide,lahar, rockfall, avalanche, mudslide, mudflow, debris
Calibration	calibration
Algorithm	algorithm
Validation	validation